TOPIC 5

Data and Data files

Concepts of data.

Data refers to raw, unprocessed facts and figures that on their own have no meaning. Examples: numbers, names, measurements, images, sounds, survey, responses, temperatures etc

Characteristics of Data.

1. Raw; Collected in its original form before processing.
2. Unorganized.
3. Can be qualitative or quantitative.
4. Can come from many sources.
5. Needs processing to become information.

Bit

- A bit is the smallest unit of data a computer can store or process
- The word bit comes from Binary Digit.
- A bit can only have two values:
  - 0 (Off/false)
  - 1 (On/true).

Representation of Data using Bits

- All types of data in IT and Statistics are represented using bits:
  - Numbers.
  - Letters.
  - Images.
  - video.
  - Sound.

Example.

The Letter A is stored as 01000001 in binary.

By

Byte

- A byte is a unit of digital data used in Computers.
- It is made up of 8 bits.
- A byte is the Standard unit for representing a Single Character such as:
  - a letter (A, B, C).
  - A digit (1, 2, 3).
  - A symbol @; % ~~#~~.

Uses of a byte

- Measure data size (files, Memory).
- Store characters in text.
- Represent small integers.
- Build larger units of data.

Data Types

Data types refers to the classification of data based on the kind of values it holds and how it can be processed by a Computer or Statistical program. They help in:

- ✓. Data Storage.
- ✓. Data analysis.
- ✓. Choosing the right Statistical Methods
- ✓. Error Checking in software.

(A) Numerical (Quantitative) Data Types.

- These represent numbers that can be Measured or Counted.

1. Integer.

- Whole numbers (no decimal point).
- Example: 5, -10, 250.
- Used to count things (eg Number of students).

2. Float / Real / Double
- Numbers with decimals.
- Examples: 3.14, 2.5, -0.01.
- Used for measurements (height, weight, temperature).

(B) Character and Text (String) Data Types

1. Character.
- A single symbol or letter.
- Example: 'A' '7' '@'

2. String / Text.
Sequence of characters (words, sentences).
Example: "Kenya", "female," "BSc Statistics"
Used for names, categories, labels, addresses etc.

(C) Logical / Boolean Data Type.
- Holds only two values: TRUE or FALSE.
- Represented as 1 or 0 internally (bit-level).
- Used in decision-making and conditions (eg Pass/fail, yes/NO).

(D) Date and Time Data Types.
used to represent
- Dates (eg 2025-12-10).
- Time (eg 30:00)
- Datetime (combination)

- Important in time series analysis in Statistics

2.
3
4

Constructing Random and Sequential Data files

- A data file is a collection of related records stored on a storage device (hard disk, USB, etc)
- In IT for statistics, data files are used to store data sets for processing by software like SPSS, R, Excel, etc.

There are two main ways of organizing data in files:

i) sequential files.
ii) Random (Direct Access) files

1) Sequential Data files

A sequential file is a file where records are stored and accessed in a specific order, usually according to the time they were entered or a key field.

characteristics

- Records stored in sequence (eg alphabetical order, date order).
- Access is linear (one by one)
- Efficient for processing large files.
- Slower when searching for a specific ~~order~~ record.

Constructing Sequential files.

Steps

Define the structure

2. Collect and enter data.
3. Store the record in the file in the same order.
4. When adding new records:
   • Often requires rewriting the entire file to keep the sequence

Advantages

- Simple to Create and Manage
- Good for batch processing (eg. Payroll, exam results).
- Uses less storage and faster to read large files sequentially

Disadvantages

- Slow when searching for specific records
- Hard to insert or delete records
- Must start reading from the beginning.

(ii) Random (Direct Access) Data files

- A random (direct-access) file is a file where records are stored in such a way that they can be accessed directly, without reading the previous records.
- A record's location is determined by a key field using a hashing algorithm or fixed-length slots.

Characteristics

- Records can be accessed instantly.
- uses a key (eg ID number) to locate the record.
- More flexible and faster for updates and searches.
- Requires more storage space than

sequential files.

Constructing Random files

steps

1. Define the key field.
2. Use a hashing function.
3. Store each record.
4. If two records land in the same slot (Collision):
   - Use collision handling Methods:
     - v. Linear probing (next available slot).
     - v. Chaining (Link Records together)

Advantages

- very fast access and retrieval.
- Easy to update, insert or delete records.
- Ideal for applications requiring frequent searches.

Disadvantages

- v. More complex to design.
- v. Uses More Storage space.
- v. Collisions May occur and Must be handled.